UCRL-JC-123863
PREPRINT

Mining Scientific Data Archives through Metadata Generation

R. Springmeyer
N. Werner
J. Long

This paper was prepared for submittal to the
First Institute for Electrical and Electronics Engineers Metadata Conference
Silver Spring, MD
April 16-18, 1996

April 1997

Lawrence Livermore National Laboratory

Mining Scientific Data Archives through Metadata Generation

Rebecca Springmeyer, Nancy Werner, and Jeffery Long, Lawrence Livermore National Laboratory

First IEEE Metadata Conference, April 16-18, 1996, NOAA Auditorium, Silver Spring, Maryland

ABSTRACT

Data analysis and management tools typically have not supported the documenting of data, so scientists must manually maintain all information pertaining to the context and history of their work. This metadata is critical to the effective retrieval and use of the masses of archived data, yet little of it exists on-line or in an accessible format. The exploration of archived legacy data typically proceeds as a laborious process in which people use commands to navigate through file structures on several machines, examining files with various applications. Adequate data exploration and mining support requires replacing this file-at-a-time approach with a model that represents data as collections of inter-related objects. The tools that support this model must focus attention on data while hiding the complexity of the computational environment.

We have addressed this problem by developing a tool for exploring large amounts of data in UNIX directories via the automatic generation of metadata summaries. In this paper we describe our model for metadata summaries of collections and our Data Miner tool for interactively traversing directories and automatically generating metadata that serves as a quick overview and index to the archived data. The summaries include thumbnail images as well as links to the data, related directories, and other metadata. Users may personalize the metadata by adding a title and abstract to the summary, which is presented as an HTML page viewed with a World Wide Web browser. We have designed summaries for three types of collections of data: the contents of a single directory; virtual directories that represent relationships between files scattered across physical locations; and groups of related calculation files. By focusing on the scientists' view of the data mining task, we have developed techniques that assist in the "detective work" of mining without requiring knowledge of mundane details about formats and commands. We present our experiences in working with scientists to design the Data Miner and associated summaries, and discuss feedback from scientists in using this tool to mine their archived data.

INTRODUCTION

The typical environment of most scientific researchers includes multiple hardware platforms, a large collection of eclectic software applications, data stored on many devices in many formats, and little standard metadata or accessible documentation about the data. Data analysis tools typically have not supported the creation of metadata, so each scientist has devised a different method to capture the relevant metadata. This situation often results in such diverse and scattered collections of data that much of it lies unused because the owner does not recall its existence, location, or content. This characterization holds true for many scientists who use the Livermore Computing facilities to produce large, complex data sets that are stored on a variety of machines and in mass storage systems. There currently does not exist an integrated set of tools that provides adequate support for accessing, organizing, and annotating these data sets and the associated collections of experimental data and documentation. The problem exists even on individual workstations, as illustrated by computer users who browse directories of files and ask "What's in here and why did I keep it?"

One way to answer this question is to check the system metadata, either through commands such as "ls" and "file" or with a file browser. This will reveal who owns the file, when it was created or modified, the size of the file, and perhaps the generating application. The contents of the file remains a mystery, unless it is an image file and the browser supports icons. In the case of exploring legacy data in archives that contain disparate kinds of data, scientists are faced with a laborious process in which they move around in file structures on several machines, examining individual files with many different applications. In order to better support data exploration and mining, scientists require a flexible, organized presentation of metadata that supports more efficient searching and browsing of hierarchies of data. This paper describes a model and data mining tool for automating a laborious part of scientists' existing data exploration and mining activity. The resulting summaries serve not only as immediate guides to the data, but as documentation that can be shared with others.

RELATED WORK

We define scientific data mining as the process of surveying and interpreting archived scientific data so that it can be applied to a current analysis problem. In the case of our collaborating scientists, the archived calculational data is not contained in traditional databases. Intelligent interfaces have been developed to explore traditional databases of objects and attributes, but such data differs from scientific data in format and scope [4]. Furthermore, artificial intelligence techniques for database mining, such as clustering methods for discovering interesting relationships and characteristics, are not directly applicable. The goals of scientific data mining are to find and survey relevant input parameters, restart files, and assorted types of data, as part of the information-gathering stage of a project. Areas of work more directly related to this topic include file browsers, web-based metadata systems, metadata standards, and systems for browsing document collections.

Directory Browsers and Interactive Metadata Creation

File browsers present files in text lists or with icons which typically represent the generating application rather than the content of the file, unless it is an image for which a thumbnail can be generated. In order to get a feel for the actual contents of the files in a directory, users must examine each file individually by launching applications. Using this file-at-a-time model, many systems provide users with the ability to retrieve system-level metadata. For example, the Macintosh Finder provides a Get Info utility for files that displays system information and one editable comments field. The fields included in the display are somewhat tailored to the types of object, such as hard disk, text file, or audio CD file. The Silicon Graphics Directory View Window also displays metadata about directories and includes a Get Info operation. Features include a resizable display with a choice of icons or listing information, search filters, a shelf for frequently-accessed files, and thumbnails for graphics. Files can be sorted by name, date, size, or type, as with the Macintosh.

Interactive metadata creation is supported by many programs, including popular presentation and document processing programs. The inclusion of document imaging and paper-based input for metadata creation has been effectively demonstrated by the Protofoil system for storing, retrieving, and manipulating documents [9]. The system's search results can be presented in five views: an array of thumbnails, lists of attributes, document category groupings, document clusters, and snippets based on keywords.

Taking file browsing a step further, some systems have moved into the realm of three-dimensional visualization, including SGI's File System Navigator [12] and the Hyper-G Internet information management system [1]. Xerox PARC researchers have developed several techniques for information visualization which allow users to interactively explore workspaces. Their Information Visualizer [2] provides three-dimensional representations including a perspective wall [8] and cone tree [10]. Advancing the visualization level with such techniques clearly assists users in navigating through a workspace. For data mining purposes, in addition to navigation, users require detailed, content-based summaries of the objects in the collection. In our own previous work, we established that visualization tools do not solve the larger scientific data analysis problem [11]. Similarly, three dimensional visualizations for browsing workspaces do not solve the larger scientific data mining problem.

Instead of asking users to become familiar with a different browser for each machine, the Intelligent Archive (IA) project at Lawrence Livermore National Laboratory (LLNL) developed a graphics file browser and FTP client for transferring files and organizing directories on a variety of local and remote hosts. LLNL XDIR displays windows representing any combination of local and remote machines. Users initiate FTP transfers by dragging and dropping files between windows. Each window has controls for setting modes and invoking operations, such as viewing graphics or text files, or launching applications. Users see the same interface on all of the machines and have a graphical representation of their files, including those stored in archival storage systems. LLNL XDIR is part of a suite of IA tools for accessing and organizing information [6], including an interactive metadata editor and the prototype data mining tool presented in this paper.

Web-based Metadata Systems

There are many examples of web-based metadata systems for scientific data retrieval, such as that of the Defense Nuclear Agency, which has a major effort in progress to archive their data [5]. The Jet Propulsion Laboratory is developing a data retrieval system (DARE) for them that includes browse representations and access algorithms for DNA data. The DARE effort demonstrates how web-based access to metadata can provide scientists with a comprehensive resource that documents important data. Similarly, the Environmental Resources Information Network [3] has set up information retrieval services with a hypermedia interface and searching via Wide Area Information Servers (WAIS). They separate metadata objectives into two classes: discovery of data sets that otherwise would have gone unnoticed, and documenting the content, quality, and features of a data set. By separating discovery from in-depth documentation, they hope to encourage the generation of metadata with their minimalist approach of about 20 core fields. This approach is shared by the Nuclear Weapons Information Group (NWIG), which developed a metadata format for standardized catalogs to describe scientific data associated with the US Department of Energy Weapons Complex [7].

The metadata model of this paper concerns the case in which scientists are browsing directories of unknown files for which they want to generate content-based metadata on the fly. Thus it differs from many of the large metadata efforts for formally cataloging data to be shared among large groups of scientists, although it also provides a web-based solution to retrieving metadata.

APPROACH

Our approach has been to work closely with end users during the development of our model and the iterative design and implementation of our prototype data mining tool. The goal is to help scientists focus more on data comprehension and less on the mechanics of searching for, transferring, and examining individual files. We have developed a model that defines a common format and protocol for creating and using browse representations -- summary information about an object, such as a data file, or collections of information, such as a directory or a hardcopy notebook. We have defined three classes of browse representations for exploring and mining archives of scientific data:

- content-based summaries of files in a directory,
- summaries of "virtual directories" that combine physically diverse files into one "logical" collection. This allows a set of results files to be included in the summaries for "virtual directories" of several different projects. Another example would be to generate one summary of all files ending in a particular extension, such as .fm, .rgb, or .c, across any number of directories, and
- summaries of simulation results (conceptually linking input files, results files, plots, and edits).

As depicted in Figure 1, each one of these summaries points off to the data itself and to related metadata, including fields such as owner, abstract, thumbnail images, keywords, date, and pointers to related information. This metadata is a subset of the NWIG metadata [7]. Tools for further investigation, such as visualization and scientific data analysis tools, are integrated into the scientific data mining process via web browser mechanisms for launching applications.

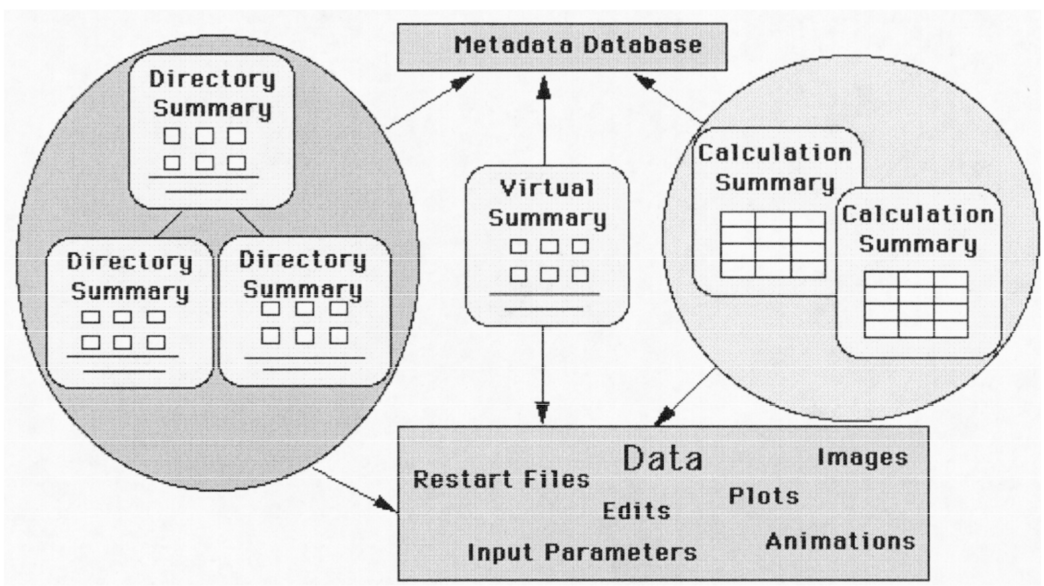


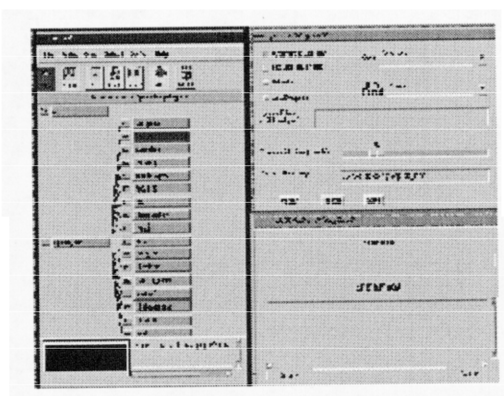
Figure 1. All three classes of browse representations point back to data as well as further metadata, presenting users with a web of easily-accessed resources.

One goal of this work is to build the architecture to be flexible and adaptable so that it will be easy to add new browse representations and thus take advantage of the functionality of the mining, viewing, and searching capabilities available through the data mining tool and an associated database of metadata. Rather than developing one mining tool per kind of data, which is essentially the current situation with custom tools, we are applying this framework consistently to several kinds of metadata. Furthermore, once users learn how to view one type of summary with our tool, they can apply it to many other kinds of data.

IMPLEMENTATION

Based on discussions and continuing feedback from end user physicists, we designed and implemented a prototype Data Miner tool. It provides a graphical user interface for navigating directories, specifying output parameters and styles, and creating, and annotating summaries. The specified summaries are created by a perl script that generates HTML with links to the data and associated NWIG-style metadata. Thumbnail graphics are created for each image file, otherwise an icon representative of the file type is used. We are currently investigating content-based icons for non-image data.

The data mining tool was designed with reusable, general components which can be applied to many scientific disciplines. We use public domain World Wide Web and graphics technology. And we are leveraging the Intelligent Archive MetaMaker tool for automatic generation of summaries about simulation runs. Data Miner generates summary metadata automatically, although we could also launch an interactive metadata editing tool for users who want to edit the standard metadata for a particular object in the collection. The following four figures and associated tables summarize the features of Data Miner. They illustrate the user interface and three examples of browse representations for summarizing collections of objects.

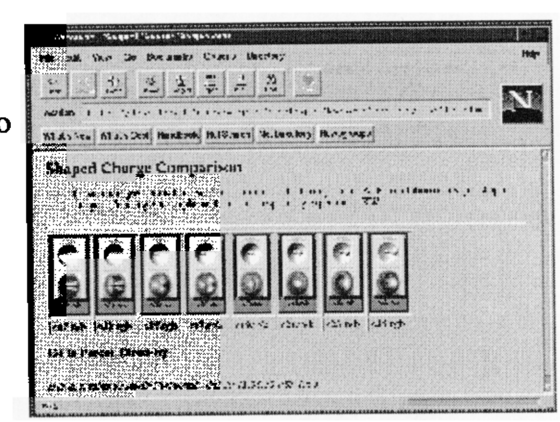


DataMiner Graphical User Interface

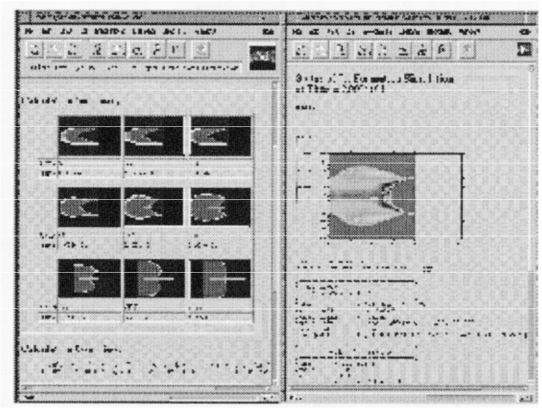
- Shows directory structures as trees, with user-specified root directory.
- Users control summary style, thumbnail size, and file protection level.
- Highlights directories with existing summaries; highlights directories with files found by using wildcard search.

Summary of Directory Contents

- Shows content of directory "at-a-glance", with links to raw data and metadata.
- Can include a user-specified title and abstract.
- Generated in HTML format for Web browsing.
- Several styles and formats are available.

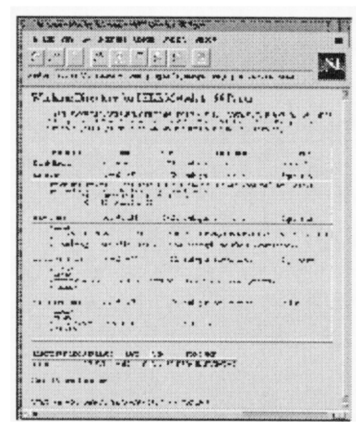


Summary of a Calculation



- Collects related calculation pieces into one package.
- Calculation summaries link to individual timestep pages.
- Timestep pages include graphics, text, and pointers to the raw data.
- Can be adapted to create tailored summaries for each simulation code.

Verbose Summary of Directory Contents



- Provides added metadata, especially for non-image files, that includes date, size, file permissions, and file type.
- Text files are described with additional lines of information, depending on the type of text file. E.g. first four lines of text files, first four lines of email after header, first four lines of source code.
- Names of files and directories are links to the files or other Data Miner summaries.

This approach has allowed us to explore several kinds of browse representations with minimal development, while taking advantage of existing tools. These include WWW browsers, browser remote control scripts, and the ImageMagick suite of image tools. The graphical user interface depends on X Windows and Motif. The perl script can be run on its own to generate metadata summaries without using a graphical user interface. This mode of operation can be used to generate the summaries as a background process, so that metadata can be created or updated for a directory structure in off-hours.

USER FEEDBACK

The metadata representations described here were developed over the course of several months, based on our interactions with scientists who explained to us their current approach to exploring and mining their archives of data files. In a series of individual interviews with several physicists, we observed how people are currently mining their data and initiated discussions about how on-line tools for creating, searching, and browsing metadata might help. We began those discussions by showing paper sketches, moving on to hand-made HTML pages, and finally presenting an initial prototype tool with a simple menu interface. Throughout this process, we refined the browse representations based on their feedback.

One major lesson was that metadata details should be hidden behind links, in favor of a "clean and simple" style of summary. For cases where more detail would be welcome, such as in browsing email archives, we developed a "verbose" style of summary. Another major request was to tune the performance of the Data Miner tool to generate summaries quickly enough to be useful for interactive browsing. Users preferred to sacrifice consistency in size of thumbnails in favor of quicker access, for example.

One of the more productive ideas came out of a series of sessions where a scientist ran the Data Miner on his files and talked about how it might be useful to create collections of files spread across several directories. This resulted in the creation of the Virtual Summary representation, which we are continuing to refine.

Another particularly well-received idea has been the creation of calculation summaries. As scientists have been shown the initial web-based prototype, they have immediately begun to provide feedback on what types of information to include and have stated strong support for this style of browse representation. Several code groups have expressed interest in adapting the MetaMaker summaries to their code and generating tailored representations for their users.

As we give demonstrations of Data Miner to physicists and computer scientists, we continue to gather feedback about how summaries can be tailored to help people with different kinds of tasks. These tasks have included mining calculation data, browsing archives of photographs of experiments, cleaning up directories of HTML files, or investigating source code directories. The key criteria for success have been simplicity, speed, and how well the summaries combine diverse pieces of data and information into one convenient package.

CONCLUSIONS

Scientists need tools for accessing and organizing many kinds of data, and they have neither the time nor the inclination to interactively document collections that they have either created or inherited. Tools are needed to automatically generate metadata in a form that summarizes the data and allows easy access to the raw data, in a way that inspires end users to augment the metadata with their own content-based metadata. Our model and the Data Miner prototype tool have succeeded so far in capturing the attention of our collaborating scientists and in leading to a revised model and user interface. Future work will include exploring methods for improved representation of non-image data, expanded browse representations, and creating an expanded virtual directory capability.

ACKNOWLEDGMENTS

We acknowledge Carol Hunter, Bruce Lownsbery, and Dick Watson for their input throughout the development of the model and the Data Miner tool. We acknowledge Neale Smith as the author of LLNL XDIR and developer of the IA metadata editing tool QuickNotes. We are especially grateful to Kris Winer for his substantial feedback as a collaborating scientist, particularly in relation to virtual directories. We thank Roger Crawfis and Neale Smith for useful comments on drafts of this paper. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract number W-7405-ENG-48, with specific support from an internal LDRD grant.

REFERENCES

[1] Keith Andrews, **Visualizing Cyberspace: Information Visualization in the Harmony Internet Browser**, Proceedings of Information Visualization '95, October 30-31, 1995, Atlanta, GA, pages 97-104.

[2] Stuart K. Card, George G. Robertson, and Jock D. Mackinlay, **The Information Visualizer, An Information Workspace**, In Proceedings of CHI '91: Human Factors in Computing Systems, May 1991, New York, pages 181-188.

[3] David Crossley, **WAIS through the Web - Discovering Environmental Information**, Proceedings of the Second International WWW Conference (WWW Fall 94) Mosaic and the Web, Chicago, IL, October 17-20, 1994. Available at http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/crossley/paper.html.

[4] Jade Goldstein and Steven F. Roth, **Using Aggregation and Dynamic Queries for Exploring Large Data Sets**, Proceedings of CHI '94 Conference on Human Factors in Computing Systems, April 24-28, 1994, pages 23-29.

[5] J. Hyon, R. Borgen, **Data Archival and Retrieval Enhancement (DARE) Metadata Modeling and Representation**, First IEEE Metadata Conference, April 16-18, 1996, NOAA Auditorium, Silver Spring, Maryland.

[6] **Intelligent Archive: Integrated Data Access and Organization for Scientists**, LLNL Technical Report, UCRL-TB-118571, November 1995. Available at http://www.llnl.gov/ia/.

[7] Bruce Lownsbery and Helen Newton, **The Key to Enduring Access: Multi-organizational Collaboration on the Development of Metadata for Use in Archiving Nuclear Weapons Data**, First IEEE Metadata Conference, April 16-18, 1996, NOAA Auditorium, Silver Spring, Maryland

[8] Jock D. Mackinlay, George G. Robertson, and Stuart K. Card, **The Perspective Wall: Detail and Context Smoothly Integrated**, In Proceedings of CHI '91: Human Factors in Computing Systems, May 1991, New York, pages 173-179.

[9] Ramana Rao, Stuart K. Card, Walter Johnson, Leigh Klotz, and Randall H. Trigg, **Protofoil: Storing and Finding the Information Worker's Paper Documents in an Electronic File Cabinet,** Proceedings of the CHI '94 Conference on Human Factors in Computing Systems, April 24-28, 1994, pages 180-185.

[10] George G. Robertson, Jock D. Mackinlay, and Stuart K. Card, **Cone Trees: Animated 3D Visualizations of Hierarchical Information**, In Proceedings of CHI '91: Human Factors in Computing Systems, May 1991, New York, pages 189-194.

[11] Rebecca R. Springmeyer, Meera M. Blattner, Nelson L. Max, **A Characterization of the Scientific Data Analysis Process**, Proceedings of IEEE Visualization '92, October 1992, pp. 235-242.

[12] Joel Tesler and Steve Strasnick, **FSN: The 3D File System Navigator**, Silicon Graphics, Inc., Mountain View, CA, 1992.